EFFICIENT GMM ESTIMATION WITH INCOMPLETE DATA

Chris Muris*

Abstract—In the standard missing data model, data are either complete or completely missing. However, applied researchers face situations with an arbitrary number of strata of incompleteness. Examples include unbalanced panels and instrumental variables settings where some observations are missing some instruments. I propose a model for settings where observations may be incomplete, with an arbitrary number of strata of incompleteness. I derive a set of moment conditions that generalizes those in Graham's (2011) standard missing data setup. I derive the associated efficiency bound and propose efficient estimators. Identification can be achieved even if it fails in each stratum of incompleteness.

I. Introduction

INCOMPLETE data, where some observations are missing some or all variables, is prevalent in empirical research in economics. For example, Abrevaya and Donald (2017) find that incomplete data occur in at least 40% of the publications in top economics journals. In 70% of these cases, all incomplete observations are discarded, and the analysis is then carried out with the resulting complete subsample. This strategy fails to use all the information in the data, since incomplete observations typically have some information about model parameters. This paper shows how to use this information.

I provide a general framework for efficient parameter estimation using incomplete data. To see why a serious treatment of incomplete observations can be useful, consider a linear instrumental variables model with two endogenous variables, $X = (X_1, X_2)$, and two instruments, W_1 and W_2. The parameter vector β_0 is defined through the moment conditions:

$$E\begin{pmatrix} W_1 (y - X\beta_0) \\ W_2 (y - X\beta_0) \end{pmatrix} = 0. \tag{1}$$

Now consider a setting where either instrument can be unavailable. This implies the existence of three strata based on data availability. In the first stratum, both instruments W_1 and W_2 are observed; in stratum 2, only the instrument W_1 is observed; in stratum 3, only W_2 is observed. Although the parameter is not identified in stratum 2, the moment condition $E[W_1 (y - X\beta_0)] = 0$ still contains information on β_0. The same is true for stratum 3 through $E[W_2 (y - X\beta_0)] = 0$. This paper provides an efficient estimator that uses information from all strata. The approach is general: it allows for an arbitrary number of strata of incompleteness and an arbitrary set of nonlinear moment conditions.

Currently available procedures for dealing with incomplete data can be classified into three categories. The first approach is to classify data (or equivalently, moments) in terms of binary missingness, that is, as either complete or completely missing. A second approach is to provide tools that work only in specific applications. The third approach is to impute the incomplete data.

The approach proposed here is distinct from all of those. I focus on incomplete data that may be partially missing. Whereas binary missingness implies the existence of exactly two strata, I allow for an arbitrary finite number of strata based on the availability of each moment. My approach accommodates any model that can be expressed in terms of moment conditions. In contrast, model-specific solutions for one type of application may not be useful for another. My approach does not require imputation. Imputation approaches have the obvious drawback that they are inconsistent if the imputation model is misspecified. My approach is consistent in part because it does not use an imputation model.[1]

This paper has three methodological contributions. First, I generalize the moment conditions established by Graham (2011) for the binary missingness case to the general incompleteness case. The resulting set of moment conditions consists of one set of Graham's moment conditions for each stratum of incompleteness.

Second, I derive the efficiency bound associated with the complete set of moment conditions and propose an estimator that attains that bound. I provide conditions under which the estimator is consistent and asymptotically normal. A simulation study (appendix D) shows that the efficiency gain from using incomplete observations can be substantial. I also propose and analyze a doubly robust estimator.

Third, I show that the parameters of interest can be identified by using all the available data, even if identification does not hold in every stratum. As an example, consider a linear IV model with two endogenous variables and two instruments, where the instruments are never observed in the same stratum, but each instrument is available from a different stratum. In this setting, one can still identify the regression parameters.

The results in this paper can also be applied to (dynamic) panel data models; equation systems where some equations have missing dependent variables for some observations; triangular simultaneous systems with some endogenous

Received for publication July 20, 2016. Revision accepted for publication March 11, 2019. Editor: Bryan S. Graham.

*Muris: University of Bristol.

I am grateful to Ramon van den Akker, Richard Blundell, Otilia Boldea, Irene Botosaru, Pedro Duarte Bom, Katherine Carman, Matias Cattaneo, Miguel Atanasio Carvalho, Bryan Graham, Hide Ichimura, Toru Kitagawa, Tobias Klein, Andrea Krajina, Jan Magnus, Bertrand Melenberg, David Pacini, Krishna Pendakur, Franco Peracchi, Pedro Raposo, Sami Stouli, Thomas Vigie, Bas Werker, and Frank Windmeijer for encouraging and insightful discussions. I also thank the seminar participants at Tilburg University, University of Bristol, Institute of Advanced Studies Vienna, Simon Fraser University, Monash University, Victoria University, and the Bristol Econometrics Study Group. I gratefully acknowledge financial support from the Social Sciences and Humanities Research Council through Insight Development Grant 430-2015-00073.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/rest_a_00836.

[1] However, I show in section VC that imputation may be useful if used in the context of doubly robust estimation.

The Review of Economics and Statistics, July 2020, 102(3): 518–530

https://doi.org/10.1162/rest_a_00836

explanatory variables missing for some observations; and general nonlinear instrumental variables models. In appendix D1, I analyze a dynamic panel data model where cross-section units may miss observations in any combination of time periods.

The paper is organized as follows. Section II provides a literature review. Section III describes the model. Section IV presents the efficiency bound results, and section V presents an efficient IPW estimator and a locally efficient doubly robust estimator. Section VI contains an empirical illustration. The appendix contains proofs, additional examples, additional material for the empirical application, and a simulation study.[2]

II. Related Literature

The literature on missing and incomplete data is vast. I discuss the relevant literature in three strands. The first strand considers efficient estimation under the assumption that every observation is either complete or completely missing. The second strand of literature considers estimation with incomplete observations for specific models. The third strand of literature augments incomplete observations using imputation. To the best of my knowledge, my paper is the first that provides a general framework for efficient estimation with incomplete observations without using imputation.

To facilitate this discussion, let p be the number of elements in a moment vector ψ, and let D be a $p \times p$ diagonal matrix with 1 on the main diagonal if a moment is observed and 0 otherwise. The incomplete data indicator D defines the strata of incompleteness in the data, and the vector $D\psi$ gives the observed elements of ψ. In the linear IV example given above, $p = 2$, and the 2×2 matrix D can take three values corresponding to 0s and 1s on the main diagonal. The three values that D can take on correspond to the three strata of data incompleteness. We say a parameter is identified in a stratum D if $D\psi$ contains enough information to identify the parameter. In the example above, the only stratum in which the parameter is identified is stratum 1 (for which $D = I_2$).

A. *Strand 1: Binary Missingness*

There is an extensive literature on missing data models in which each observation contributes either to all or to none of the sample moments (i.e., the missing data indicator is a binary variable). This literature typically employs the missing-at-random (MAR) assumption. I call models including a MAR assumption the *MAR setup* (as in Graham, 2011).

The literature on the MAR setup was initiated by Robins, Rotnitzky, and Zhao (1994), who propose an augmented inverse propensity score weighting (AIPW) procedure. An overview of the AIPW literature in statistics can be found in Tsiatis (2006). Chen, Hong, and Tarozzi (2008) derive the efficiency bound for nonlinear and possibly overidentified models and propose an efficient estimator for the parameters in the MAR setup that is not based on inverse propensity score weighting (IPW). An important result in this literature is that estimating the propensity score is more efficient than using the true value of the propensity score (the IPW paradox; Hirano, Imbens, & Ridder, 2003; Wooldridge, 2007; Prokhorov & Schmidt, 2009).

Two contributions from this literature that are especially relevant for the discussion in this paper are Graham (2011) and Cattaneo (2010). Graham (2011) shows that in a MAR setup with binary missingness (just two strata for D), the efficiency bound is equivalent to the efficiency bound for the inverse weighted moment conditions of the original (complete data) model, plus a set of conditional moment conditions that captures all the information from the MAR assumption. I generalize the moment conditions established by Graham (2011) for the binary missingness case to the general incompleteness case with J strata.

Cattaneo (2010) considers the efficient estimation of multivalued treatment effects.[3] His model is similar to mine, with incompleteness taking the form of missing dependent variables. With multivalued treatment effects, this incompleteness implies as many strata as there are levels of treatment. Cattaneo shows how to optimally combine the information from the different values of the treatment, but his approach requires that the parameter vector is identified in each stratum. Consequently, his approach cannot be used for the linear IV example given above. I provide sufficient conditions for an optimal estimator when the parameter vector is identified in just one stratum. Further, I provide special cases where identification is not required in any stratum. More details on this comparison can be found in appendix B. A related contribution is in Chaudhuri and Guilkey (2016).

B. *Strand 2: Model-Specific Solutions*

Several papers consider specific GMM settings or specific incomplete data patterns. For example, Abrevaya and Donald (2017) consider the linear regression model. Model-specific solutions are also available for the instrumental variable model with incomplete sets of instruments. The problem of partially missing instruments is common (see Angrist, Lavy, & Schlosser, 2010). Instrumental variables estimation with missing instruments is discussed in Mogstad and Wiswall (2012), who consider a setting with a single instrument that is missing for a subsample of the observations. Abrevaya and Donald (2011) also consider the missing instrument model.

Chen, Yi, and Cook (2010) provide an estimator for the parameters in a static panel data model. Verbeek and Nijman (1992) also consider the static model and propose to use

[2] The appendixes are part of the supplementary material, accessible through the journal's website.

[3] The relationship between the multivalued treatment effect setting in Cattaneo (2010) and the incomplete data setting here is described in more detail in appendix B.

the different missing data patterns to test for selectivity bias. Hirano et al. (2001) consider a panel data model with three strata of incompleteness.[4] Abrevaya (2019) shows that the explanatory variables in the static model have information even when the associated dependent variable is unavailable.

The linear dynamic panel data model with attrition has recently been considered by Pacini and Windmeijer (2015; see also their references). Pacini and Windmeijer (2015) show that nonlinear, previously not considered moment conditions are informative when data from some time periods are unavailable.

My approach accommodates any model that can be expressed in terms of moment conditions and allows for any structure of incompleteness. In contrast, model-specific solutions restrict the structure of incompleteness, and solutions for one type of application may not be useful for another.

C. *Strand 3: Imputation*

A substantial literature considers augmenting incomplete observations by imputing the unavailable components. A leading example is the linear regression model with missing covariates. Using variables that are always observed, an imputation model can be estimated using the complete observations, and it can then be used to fill in the incomplete observations. Early contributions to the econometric literature on this topic can be found in Dagenais (1973) and Gourieroux and Monfort (1981). To retain consistency, these approaches require a correctly specified imputation model. Such an assumption is not maintained in the model that I consider. A more recent contribution by Dardanoni, Modica, and Peracchi (2011) shows that efficiency gains can be obtained if one is willing to sacrifice consistency.

In the context of the linear IV example above, imputation would apply to missing instruments. If the imputation were correctly specified, then imputation would not result in bias and would improve the efficiency of the estimator. However, under misspecification, the resulting estimator would typically be biased. My approach does not require imputation: I propose an inverse propensity score weighting estimator that is consistent.[5]

III. Model

This section formalizes the notion of incomplete data in this paper and introduces identification and sampling assumptions that are used throughout the paper.

A. *Incomplete Data*

The incomplete data framework starts from moment conditions for complete data. Let $Z = \left(Y_1', X'\right)$ be a random vector of data, let β be an unknown parameter vector of size $K \times 1$, and let $\psi(Z, \beta)$ be a $p \times 1$ vector of moment functions, with $p \geq K$.[6] The true value of the parameter, $\beta_0 \in \mathscr{B} \subset \mathbb{R}^K$, is defined by assumption 1.

Assumption 1. $E(\psi(Z, \beta)) = 0 \Leftrightarrow \beta = \beta_0$.

In this paper, not all elements of the vector $\psi(Z, \beta)$ are always observable. To model this, let D be an incomplete data indicator with $J + 1$ outcomes, or incomplete data patterns, $\{d_1, \ldots, d_{J+1}\}$. Every incomplete data pattern corresponds to a stratum defined by data availability. An incomplete data pattern d_j is a $p \times p$ selection matrix that selects the elements of ψ that are observable for an observation in stratum j. In other words, the researcher observes $D\psi(Z, \cdot)$. In stratum $J + 1$, none of the components of ψ are observed: $d_{J+1} = O_{p \times p}$.

The following three examples illustrate the setup. (Additional examples can be found in appendix C.)

Example 1: Linear IV. Consider a linear instrumental variables model with a dependent variable y, two endogenous variables $X = (X_1, X_2)$, and two instruments $W = (W_1, W_2)$. Set $Z = (y, X, W)$, and define the moment function,

$$\psi(Z, \beta) = \begin{bmatrix} W_1 (y - X\beta) \\ W_2 (y - X\beta) \end{bmatrix},$$

so that the parameter vector β_0 is defined through the moment condition $E(\psi(Z, \beta_0)) = 0$. The incomplete data indicator takes one of $J + 1 = 4$ values:

$$D \in \left\{ d_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, d_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, d_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, d_4 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right\}.$$

For d_1, this corresponds to observing all variables,

$$d_1 \psi(Z, \beta) = \begin{bmatrix} W_1 (y - X\beta) \\ W_2 (y - X\beta) \end{bmatrix},$$

for any value of β. In the stratum with $D = d_2$, only the instrument W_1 is available. This corresponds to observing

$$d_2 \psi(Z, \beta) = \begin{bmatrix} W_1 (y - X\beta_0) \\ 0 \end{bmatrix}.$$

Similarly, in the stratum with $D = d_3$, only the second instrument, W_2, is observed. Finally, the stratum with $D = d_4$ corresponds to the observations for which both instruments are unavailable or for which the dependent variable or one of the regressors is not observed.

[4] An observation is either complete, subject to attrition, or part of a refreshment sample.

[5] As opposed to the IPW estimator, the doubly robust estimator in section VC does use imputation.

[6] Wherever possible, I will use the notation in Graham (2011) to facilitate a comparison with the missing at random setup in that paper.

Models with multiple, incompletely observed instruments are relevant for applied practice. Some examples include Card (1995), Rodrik, Subramanian, and Trebbi (2005), and Angrist et al. (2010). Methodological contributions include Abrevaya and Donald (2011), Mogstad and Wiswall (2012), and Feng (2018).

Example 2: Rotating dynamic panel. Consider a five-period fixed-effects autoregressive distributed lag model with regression equation

$$Y_{it} = \alpha_i + \rho Y_{i,t-1} + X_{it}\beta_1 + X_{i,t-1}\beta_2 + u_{it},\ t = 1, \ldots, 5. \tag{2}$$

Because of the presence of the fixed-effects α_i, estimation of the parameters $\theta = (\rho, \beta_1, \beta_2)$ is based on the regression equation in first differences:

$$\Delta Y_{it} = \rho \Delta Y_{i,t-1} + \Delta X_{it}\beta_1 + \Delta X_{i,t-1}\beta_2 + \Delta u_{it},\quad t = 2, \ldots, 5. \tag{3}$$

In the estimation of empirical growth models and production functions, it is typically assumed that $E[\Delta u_{it} | Y_{i,t-3}, Y_{i,t-4}, X_{i,t-3}, X_{i,t-4}] = 0$.[7]

For a hypothetical unit with five time periods, we have

$$E\begin{bmatrix} Y_{i2}\Delta u_{i5} \\ X_{i2}\Delta u_{i5} \\ Y_{i1}\Delta u_{i5} \\ X_{i1}\Delta u_{i5} \\ Y_{i1}\Delta u_{i4} \\ X_{i1}\Delta u_{i4} \end{bmatrix} = 0.$$

Assume that a rotating panel is available. There are two cohorts, each providing four consecutive time periods. The first cohort enters the sample in period 1 and leaves in period 4. The second cohort enters the sample in period 2 and leaves in period 5. In that case, the incomplete data indicators are

$$\widetilde{d}_1 = \begin{bmatrix} 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&1&0 \\ 0&0&0&0&0&1 \end{bmatrix},\ \widetilde{d}_2 = \begin{bmatrix} 1&0&0&0&0&0 \\ 0&1&0&0&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \end{bmatrix}.$$

Example 6 in appendix C discusses a closely related dynamic panel model with a more complex pattern of missingness. Such examples are abundant in empirical work (see the applications in Arellano & Bond, 1991; Schularick & Steger, 2010; Topalova & Khandelwal, 2011; and Acemoglu et al., 2015, 2018, among many others). In section VI, I revisit the study by Topalova and Khandelwal (2011) using the methods developed in this paper.

Example 3: Panel binary choice. Consider a three-period fixed-effects logit model for the dependence of a sequence of binary outcomes $Y_i = (Y_{i1}, Y_{i2}, Y_{i3})$ on k-dimensional covariates $X_i = (X_{i1}, X_{i2}, X_{i3})$ through conditional choice probabilities:

$$P(Y_{it} = 1 | X_i, \alpha_i) = \Lambda(\alpha_i + X_{it}\beta),\ t = 1, 2, 3.$$

With complete data, estimation of the common parameters proceeds by conditional maximum likelihood, based on the conditional probability

$$P\left(Y_i = y | \sum_t y_t = c, X_i\right) = \frac{\exp\left(\sum_t y_t X_{it}\beta\right)}{\sum_{d \in B_c} \exp\left(\sum_t d_t X_{it}\beta\right)}, \tag{4}$$

where B_c is the set of all sequences d with $\sum_t d_t = c$ (see Chamberlain, 1980, and Cameron & Trivedi, 2005). Estimation of β based on equation (4) requires that all time periods are available for each cross-section unit.

I am not aware of any available estimator for β that allows for data to be incomplete at random.[8] However, the present framework easily accommodates this setting. Consider a combination of two distinct time periods, $\{(s,t) : 3 \geq t > s \geq 1\}$. The random variables $(Y_{is}, Y_{it}, X_{is}, X_{it})$ follow a two-period binary choice model, with conditional probability $P(Y_{it} = 1 | Y_{is} + Y_{it} = 1, X_i) = \Lambda((\Delta_{st}X_i)\beta)$, where $\Delta_{st}X_i = X_{it} - X_{is}$, that is, a cross-sectional logit for a subpopulation of switchers. The score is

$$E\left[A_{i,st}(\Delta_{st}X_i)(Y_{it} - \Lambda((\Delta_{st}X_i)\beta))\right] = 0, \tag{5}$$

where $A_{i,st} = 1\{Y_{is} + Y_{it} = 1\}$.

The three-period model implies three such two period models and $3k$ moment conditions:

$$E\begin{bmatrix} A_{i,12}(\Delta_{12}X_i)(Y_{i2} - \Lambda((\Delta_{12}X_i)\beta)) \\ A_{i,13}(\Delta_{13}X_i)(Y_{i3} - \Lambda((\Delta_{13}X_i)\beta)) \\ A_{i,23}(\Delta_{23}X_i)(Y_{i3} - \Lambda((\Delta_{23}X_i)\beta)) \end{bmatrix} = 0. \tag{6}$$

For a cross-section unit with complete data, the incomplete data indicator is $d_1 = I_{3k}$: all moment functions can be computed. For a cross-section unit that drops out after period 2 (attrition), $d_2 = e_{1,3} \otimes I_k$. For a cross-section unit that enters the sample in period 2, $d_3 = e_{3,3} \otimes I_k$. For a cross-section

[7] Further lags of the dependent and explanatory variables would also qualify as instruments, but are not available for any t because we are considering only five time periods. Closer lags are not valid instruments due to measurement error and endogeneity.

[8] For example, Papke and Wooldridge (2008) write: "The nonlinear models we apply are difficult to extend to unbalanced panel data—a topic for future research." Their discussion indicates Papke (2005) as an application of the methodology developed here.

unit that is not observed in period 2, $d_4 = e_{2,3} \otimes I_k$. A cross-section unit that misses more than one period has $d_5 = O_{3k}$.

The approach outlined in this example transfers to most panel models with unbalanced data. Unbalanced panels are ubiquitous in applied work across fields (see, e.g., Topalova & Khandelwal, 2011; de Loecker & Warzynski, 2012; Becker & Woessmann, 2013; Sturm & de Haan, 2015; and Yagan, 2015, among many others).

B. Identification

The following assumption guarantees identification for the incomplete data setting, given that identification holds for complete data, that is, assumption 1 holds:

Assumption 2. Every component of ψ is observable in at least one stratum, so that the matrix $\sum_{j=1}^{J} d_j$ has full rank.

Assumption 2 rules out situations in which a component of ψ is never observed. If assumption 2 fails, the analysis may proceed after removing the never-observed components from ψ (provided that assumption 1 holds for the reduced set of moment conditions).

Assumption 2 can hold *even if identification fails in every stratum*. This is an important distinction between the setup here and the multivalued treatment framework in Cattaneo (2010; see appendix B). The following examples illustrate this for two distinct cases: (a) there exists at least one stratum in which the parameters are identified, and (b) identification fails in every stratum. In case (a), standard results from the MAR setup can be applied to one of those strata, but the resulting procedure will be less efficient than the estimators proposed below. In case (b), the results proposed below are required for identification.

Example: Linear IV (continued). Recall example 1. Existing results for missing data can be used to define an estimator based on the subpopulation with both instruments observed (stratum 1, with $d_1 = I_2$). The results in my work can be applied to obtain more efficient procedures (see the analysis in section IVA).

Now consider example 4 in appendix C, which differs from example 1 because no complete observations are available. Instead, for every observation, exactly one instrument is available. This corresponds to strata 2 and 3, with

$$d_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, d_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Note that $d_2 + d_3 = I_2$, so assumption 2 is satisfied, even though identification fails in every stratum. The results below can be used to obtain a consistent and efficient estimator that deals with this problematic data setting.

Example: Rotating dynamic panel (continued). Recall example 2. For the first stratum, only two moment conditions are available to three parameters: stratum identification does not hold. Similarly, stratum identification does not hold for the second stratum. Furthermore, assumption 2 does not hold for this formulation, since $\widetilde{d_1} + \widetilde{d_2} \neq I_6$. In other words, two of the moment functions are not computable for any individual. For this reason, reduce the moment conditions to

$$\psi(Z_i, \theta) = \begin{bmatrix} Y_{i2}\Delta u_{i5} \\ X_{i2}\Delta u_{i5} \\ Y_{i1}\Delta u_{i4} \\ X_{i1}\Delta u_{i4} \end{bmatrix}$$

so that

$$D_i \in \left\{ d_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, d_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right\},$$

and assumption 2 is satisfied for the reduced set of moment conditions.

The framework in this paper can now be applied directly to estimate the parameters in the ADL model. Existing results for dynamic panel models suggest that five time periods are required for identification. However, the results that follow show that identification can be obtained using a rotating panel with four periods per individual. An efficient estimator for the parameters in that model follows immediately from the general results in this paper. This case is investigated in a simulation study in appendix D1.

Example: Panel binary choice (continued). Recall example 3. For this model, the results in this paper are not necessary for identification: the researcher could simply discard strata 2 through 5 and apply results for the standard MAR setup to the balanced subpanel (stratum 1, $d_1 = I_{3k}$). However, the efficiency gains can be substantial when the probability of missingness is large, as will be demonstrated using Monte Carlo simulations in appendix D. Similar efficiency gains may be obtained using the results in Cattaneo (2010, section 5.5).[9]

C. Sampling

The remainder of the paper analyzes efficient estimation of β_0 under the following assumptions on the sampling design and data availability.

Assumption 3. (i) Random sampling: $\{(Z_i, D_i),\ i = 1, \ldots, n\}$ is an i.i.d. sequence; (ii) the researcher observes D_i, X_i, and $D_i\psi(Z_i, \beta)$ for all $\beta \in \mathscr{B}$; (iii) missing at random: $Y_1 \perp D|X$;

[9] Strictly speaking, this would require an extension of the results in Cattaneo (2010) that allows the moment conditions to depend on the stratum. An inspection of his proofs suggests that such an extension is straightforward. See appendix B for more details on the relationship between Cattaneo's results and those in this paper.

(iv) overlap: there exists a $\kappa > 0$ such that

$$p_{j,0}(x) = P\left(D = d_j \middle| X = x\right) \geq \kappa \tag{7}$$

for all $j = 1, \ldots, J+1$ and for all $x \in \text{supp}(X)$.

This assumption generalizes the standard assumptions for missing data, in which an observation is either complete or completely missing. Assumptions 1, 2, and 3 reduce to the standard missing at random (MAR) setup if $J = 1$ and $d_1 = I_p$ (see, e.g., Graham, 2011). In what follows, I will refer to that case as "missing data" or "the standard MAR setup." One difference with the standard MAR setup is that the conditional independence assumption in part iii could be generalized to let the conditioning covariates vary by stratum, using the results in Hristache and Patilea (2016).

MAR assumption 3(iii) says that all observable data must be independent of what subset of data is available, conditional on some covariates X. This assumption is best understood in the context of an example. In the linear IV example, MAR requires instrument availability to be conditionally independent of the value of the instruments, the covariates, and the error term in the model. In the context of the panel binary choice model, it requires that the availability of data for a given cross-section unit in a certain period is independent of the fixed effect of that individual and that it is also independent of the covariates and error terms in all time periods.[10]

IV. Efficiency Bound

Assumptions 1 and 3 imply a set of conditional and unconditional moment conditions for each stratum. For each $j \in \{1, \ldots, J\}$, define the stratum indicator $s_j = 1\left\{D = d_j\right\}$. The conditional moment restrictions,

$$E\left[\frac{s_j}{p_{j,0}(X)} - 1 \middle| X\right] = 0 \text{ for all } j = 1, \ldots, J, \tag{8}$$

define the propensity scores, equation (7). In the standard MAR setup, Graham (2011) refers to such moment conditions as "auxiliary moments." Furthermore, the unconditional moment restrictions

$$E\left[\frac{s_j}{p_{j,0}(X)} d_j \psi(Z, \beta_0)\right] = 0, \; j = 1, \ldots, J, \tag{9}$$

hold. These generalize Graham's "identifying moments" to the incomplete data context. The sample analogs of moment conditions (8) and (9) can be computed with the available data (see assumption 3ii).

[10]This assumption can be weakened. The crucial assumption on independence is that the moment functions are mean-independent of the incomplete data indicator conditional on the confounders. For example, in the linear IV example, the MCAR assumption can be weakened to: "In each stratum, the observable instruments should be valid." However, with some effort, one can construct examples where identification fails under this weaker mean-independence assumption. For this reason, the stronger MAR assumption is maintained in this paper.

In what follows, denote by

$$\Gamma_0 \equiv \frac{\partial E[\psi(Z, \beta_0)]}{\partial \beta_0} \tag{10}$$

the expected derivative of the moment functions evaluated at the truth, if it exists. Also, denote by $\Sigma_0(X) \equiv \text{Var}[\psi(Z, \beta_0)|X]$ the conditional variance of the moment function.

Assumption 4. (i) The distribution of Z has known, finite support; (ii) $\mathscr{B}$ is open, and there exists a $\beta_0 \in \mathscr{B}$ and $0 < p_{j,0} < 1, \; j = 1, \ldots, J$, such that equations (8) and (9) hold; (iii) ψ is continuously differentiable on Θ for all values in the support of Z, and Γ_0 has full rank; (iv) $\Sigma_0(x)$ is invertible for all $x \in \text{supp}(X)$.

These assumptions translate the requirements for lemma 2 in Chamberlain (1987) and theorem 1 in Graham (2011) to the incomplete data setting. Below, I follow their results in constructing a semiparametric efficiency bound. Part i imposes that the data follow a multinomial distribution. The estimators I propose do not require this and still achieve the bound in the upcoming theorem 1. Remark 2 provides some additional discussion on this restriction. Part ii is not restrictive. Part iii is a strong assumption on the smoothness of the moment function. In the large sample theory developed in the remainder of this paper, this assumption is relaxed. The proposed estimators allow for nonsmooth moment conditions and still achieve the efficiency bound. Part iv requires enough variation in the conditional moments, which is readily checked in a given application.

Theorem 1 *(Efficiency Bound). If assumption 4 holds, then the information bound for any regular estimator for* β_0 *is given by*

$$I_0(\beta_0) = \Gamma_0' \left(\sum_j \left(d_j \Omega_j d_j\right)^+ \right) \Gamma_0, \tag{11}$$

where

$$\Omega_j = E\left[\frac{\Sigma_0(X)}{p_{j,0}(X)} + q_0(X) q_0'(X)\right], \tag{12}$$

$$q_0(X) = E[\psi(Z, \beta_0)|X]. \tag{13}$$

Proof. See appendix A1.

Section IVA provides an interpretation for this bound using the linear IV example. For an interpretation in the general context, recall the information bound for the binary missing data case (see, e.g., Graham, 2011),

$$I_{\text{MD}} = \Gamma_0' \Omega_1^{-1} \Gamma_0, \tag{14}$$

where Ω_1 is a stratum-specific variance as in equation (12), for the complete-data stratum with $p_{j,0} = p_0$ the standard

propensity score, and $d_1 = I_p$. First, note that the new bound in equation (11) is therefore a generalization of the bound for the MAR setup with $J = 1$, $d_1 = I_p$.

Second, note that the contribution of stratum j to the information bound is

$$I_j(\beta_0) = \Gamma_0' \left(d_j \Omega_j d_j\right)^+ \Gamma_0, \tag{15}$$

in the sense that $I_0(\beta_0) = \sum_j I_j(\beta_0)$. Compare equations (14) and (15). The new bound in equation (11) has the interpretation that it is the sum of the information in the J implied binary missing data problems.

Remark 1. The bound is reminiscent of the bound for multivalued treatment effects (see Cattaneo, 2010). Appendix B explores the relationship between the two frameworks in detail (see also section IIA). To make a comparison of the bounds, we must consider the case where $d_j = I_p$ for all j. Then the observation in section 5.5 in Cattaneo (2010) can be applied. In the framework of that section, set π equal to β_0 in this paper, and set $\beta(\pi) = (\pi, \ldots, \pi)$ so that $\partial\beta(\pi^*) = \iota_J \otimes I_p$. The equivalence of the bounds then follows immediately.

Remark 2. The bound in theorem 1 is for discrete data (assumption 4(i)). This is an approach that follows Chamberlain (1987; see also Chamberlain, 1992a, 1992b, and Graham, 2011). An alternative approach would avoid the multinomial assumption.[11] However, the bound in equation (11) can be shown to apply to arbitrary distributions.[12]

A. *Linear IV Case*

Consider example 1 (linear IV) from section III, with a set of instruments $W = (W_1, W_2)$ and an error term $u = y - X\beta_0$, $\beta_0 \in \mathbb{R}$ such that the moment conditions are given by

$$E[\psi(Z, \beta_0)] = E\begin{bmatrix} W_1 u \\ W_2 u \end{bmatrix} = 0. \tag{16}$$

Either instrument can be missing, so $J = 3$ and the incomplete data indicator has support:

$$D \in \left\{ d_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, d_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, d_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, d_4 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right\}.$$

Some additional restrictions will allow us to compare the efficiency bound in equation (11) to several estimators in common use. First, assume that $X = 1$ (i.e., incompleteness is completely at random) and that each instrument is missing with probability p, so that $p_{10} = (1-p)^2$ and $p_{20} = p_{30} = p(1-p)$. Second, unbeknown to the researcher, let $E\left(u^2 \middle| W\right) = \sigma^2$. Then

$$E\left[\psi(Z, \beta_0)\psi(Z, \beta_0)'\right] = \sigma^2 E\left[WW'\right] = \sigma^2 \Sigma_Z = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Finally, assume that the instruments are equally correlated with the endogenous variable, $E[WX] = \sigma_{xw}\iota_2$, with ι_m the unit vector of length m.

The expression for the bound now simplifies because $q_0(X) = 0$ and $\Omega_j = \frac{\sigma^2}{1-p_{j0}}\Sigma_Z$, and the expected derivative is $\Gamma_0 = -\sigma_{xw}\iota_2$. The contribution of stratum j to the information bound, equation (15), is therefore given by

$$I_j(\beta_0) = \frac{\sigma_{xw}^2}{\sigma^2}\left(1 - p_{j0}\right)\iota_2'\left(d_j \Sigma_Z d_j\right)^+ \iota_2.$$

For strata 2 and 3,

$$I_2(\beta_0) = I_3(\beta_0) = \frac{\sigma_{xw}^2}{\sigma^2}\left(1 - p(1-p)\right). \tag{17}$$

For the full data stratum,

$$I_1(\beta_0) = \frac{2\sigma_{xw}^2}{\sigma^2}\frac{(1-p)^2}{1+\rho}.$$

We can now conclude two things. First, the ratio of information in the incomplete strata 2 and 3 relative to stratum 1 is

$$\frac{I_2 + I_3}{I_1} = (1+\rho)\frac{1 - p + p^2}{1 - 2p + p^2}.$$

If $\rho = 0$, the two incomplete strata contain more information than the complete one, demonstrating that the information in the incomplete strata is not negligible.

Second, the information bound is

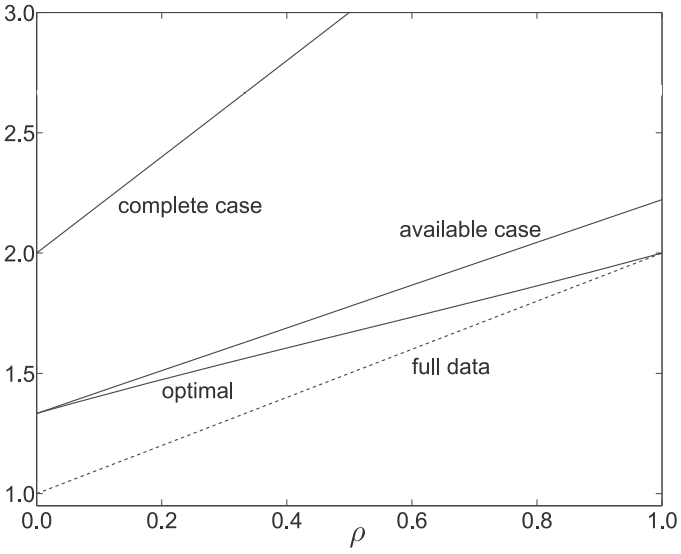
$$I_0(\beta_0) = \sum_j I_j(\beta_0) = \frac{2\sigma_{xw}^2}{\sigma^2}\left(\frac{(1-p)^2}{1+\rho} + \left(1 - p(1-p)\right)\right). \tag{18}$$

We wish to compare this bound for an optimal estimator to a few reasonable alternatives. First, the complete case estimator (CC) uses only observations with both instruments. This corresponds to the standard MAR setup and uses only stratum 1, so that $I_{CC}(\beta_0) = I_1(\beta_0)$. Second, the infeasible full data (FD) estimator has both instruments always available, with information corresponding to the standard bound for

[11] See Bickel et al. (1993), Hahn (1998), Chen et al. (2008), and Cattaneo (2010). The lack of invertibility apparent from equation (15) creates some technical difficulties in this approach.

[12] See theorem 2 in Chamberlain (1987) for the unconditional case and theorem 3 for the conditional case. Demonstrating that it can also be done for the mixed conditional/unconditional case is beyond the scope of this paper.

FIGURE 1.—INFORMATION FOR DIFFERENT SETS OF MOMENT CONDITIONS, AS A FUNCTION OF ρ, FOR $p_1 = 0.5$



equation (16), $I_{\text{FD}} = I_1(\beta_0)/(1-p)^2$. Third, the available case estimator replaces all instruments by 0s. This amounts to estimating each of the moment functions using all the observations for which that moment function is observed, with information

$$I_{\text{AC}}(\beta_0) = \frac{2\sigma_{xw}^2}{\sigma^2} \times \frac{1-p}{1+\rho}.$$

This corresponds to using the moment conditions $E[DWu] = 0$. To see this, note that

$$\mu_D \equiv E[D] = \begin{bmatrix} 1-p & 0 \\ 0 & 1-p \end{bmatrix},$$

so the available case moment conditions have derivative $-\sigma_{xw}\mu_D\iota_2 = -\sigma_{xw}(1-p)\iota_2$ and variance $\sigma^2(1-p)\Sigma_Z$.

In figure 1 we plot the asymptotic variance of the estimators, including an optimal one that achieves the efficiency bound in equation (11), as a function of ρ for $p = 0.5$.

The key aspect of this comparison is that the two instruments provide similar sources of information. Therefore, as ρ increases, two effects are expected. First, the total amount of information for β_0 decreases, so we expect the variance of all estimators to increase. Second, the amount of information on the instrument that is missing increases. Since the optimal estimator is constructed such that it efficiently exploits the correlation between the components of the moment conditions, we expect the relative performance of the optimal estimator to increase.

V. Estimation

Assume that for each stratum $j = 1, \ldots, J$, an estimator $\widehat{p}_j$ for the propensity score $p_{j,0}$ is available. Estimation of β can then be based on a matrix-weighted average of sample analogs of the feasible moment conditions, equation (9), with the propensity score estimators $\widehat{p}_j$ plugged in. The matrix weights $A_{j,n}$ are sequences of random $K \times p$ matrices, which lead to the K-dimensional sample criterion function:

$$G_n(\beta) = \sum_j A_{j,n} \frac{1}{n} \sum_{i=1}^{n} \frac{s_{ij}}{\hat{p}_j(X_i)} d_j \psi(Z_i, \beta). \tag{19}$$

The IPW estimator $\widehat{\beta}_n$ is defined as the value of β that sets that function equal to 0: $G_n(\widehat{\beta}_n) = 0$. In what follows, we will use $\|A\| = \sqrt{\text{tr}(A'A)}$ to denote the matrix norm for any matrix A. For a function $f : \mathbb{D} \to \mathbb{R}$, denote by $\|f\|_\infty$ its sup-norm $\|f\|_\infty = \sup_{x \in D} |f(x)|$.

A. *Consistency*

To establish consistency of the proposed estimator, we require some conditions on the propensity score estimators, the weight matrices $A_{j,n}$ and their limits, on the function ψ, and on the parameter space $\mathscr{B}$.

Assumption 5. For each $j = 1, \ldots, J$, the propensity score estimator is consistent:

$$\left\|\widehat{p}_j - p_{j,0}\right\|_\infty = o_p(1).$$

Assumption 5 requires the propensity score estimators to be consistent. Cattaneo (2010, appendix B) proposes a multinomial logistic series estimator that satisfies assumption 5 under mild conditions on the regressors. It can be used without modification in the present context.

Assumption 6. For each j, there exists a $K \times p$ matrix A_j such that (i) $\|A_{j,n} - A_j\| = o_p(1)$; (ii) $A_j d_j = A_j$; and (iii) $rk(A) = K$, where $A = \sum_j A_j$.

Part i is standard. Parts ii and iii are necessary for identification. They restrict the choice of limiting weights A_j to prevent underidentification. This could happen if A_j assigns zero weight to moment conditions for which the corresponding elements d_j are nonzero. If A_j is chosen as the nonzero rows of d_j, part iii reduces to assumption 2.

Assumption 7. (i) The class of functions $\{\psi(\cdot, \beta), \beta \in \mathscr{B}\}$ is Glivenko-Cantelli; (ii) $E\left[\sup_{\beta \in \mathscr{B}} \|\psi(Z, \beta)\|\right] < \infty$; (iii) $E[\psi(Z, \beta)]$ is continuous; and (iv) $\mathscr{B}$ is compact.

Part i guarantees the uniform convergence of sample averages of the original moment function ψ to its expectation. Together with part ii and the assumptions on the propensity scores and their estimators, it implies uniform convergence of the sample criterion function, equation (19). Parts iii and iv, combined with the limiting objective function having a unique 0, guarantee that the minimum of the limiting objective function is well separated (see the proof for details). These restrictions are mild. It allows for moment functions that are discontinuous, for example, a maximum score

estimator for the panel data binary choice model with attrition and refreshment.

Theorem 2 *(Consistency of IPW Estimator). Under assumptions 1, 2, 3, 5, 6, and 7,*

$$\widehat{\beta}_n \xrightarrow{p} \beta_0 \text{ as } n \to \infty.$$

B. Asymptotic Normality

We impose some additional smoothness assumptions on ψ to establish $\sqrt{n}-$asymptotic normality of the IPW estimator.

Assumption 8 (Differentiability). $E[\psi(Z, \beta)]$ is differentiable in β at β_0, and the derivative Γ_0 has full rank.

Assumption 9. For some $\delta > 0$: (i) The class of functions $\{\psi(\cdot, \beta), \|\beta - \beta_0\| < \delta\}$ is Donsker; (ii) the second moment is locally uniformly bounded:

$$E\left[\sup_{\|\beta-\beta_0\|<\delta} \|\psi(Z, \beta)\|^2\right] < \infty.$$

These assumptions are adapted from Cattaneo (2010, assumption 6). They imply stochastic equicontinuity of the criterion function. These smoothness assumptions are mild, requiring differentiability only after smoothing by taking expectations, and requiring it only at the truth. It rules in, among other things, a modification of the instrumental variable quantile regression estimator for incomplete data.

Theorem 3 *(Limiting Distribution of the IPW Estimator). Under the conditions of theorem 2 and assumptions 8 and 9 and*

$$\left\|\widehat{p}_j - p_{j,0}\right\|_\infty = o_p\left(n^{-1/4}\right),$$

then for any β_0 in the interior of $\mathscr{B}$,

$$\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathscr{N}\left(0, \left(\Gamma_0' A' V_A^{-1} A \Gamma_0\right)^{-1}\right), \tag{20}$$

where

$$V_A = \sum_j A_j \Omega_j A_j', \tag{21}$$

with Ω_j as in equation (12).

Remark 3 (Efficiency of the IPW Estimator). The asymptotic variance is minimized by setting $A_j^* = \Gamma_0'\left(d_j \Omega_j d_j\right)^+$. This resembles the usual optimal choice of weights in moment-based estimation, except for the d_j, which guarantee that only observable moment functions are selected for each stratum. Call the resulting estimator $\widehat{\beta}_n^*$. Then

$$\sqrt{n}\left(\widehat{\beta}_n^* - \beta_0\right) \xrightarrow{d} \mathscr{N}\left(0, I_0^{-1}(\beta_0)\right),$$

that is, the IPW estimator achieves the semiparametric efficiency bound derived in equation (11).

C. Doubly Robust Estimation

For the doubly robust estimator, the researcher uses possibly misspecified working models for the propensity score and the conditional expectation function.[13] Posit a working model for the propensity scores,

$$p_j(X) = \zeta_{1j}\left(h_1(X)\gamma_j\right), \; j = 1, \ldots, J, \tag{22}$$

where $h_1(X)$ is a $K_1 \times 1$ transformation of the confounders X, and the γ_j are the associated regression coefficients.

Posit a working model for the conditional expectation function,

$$q_0(X) = \zeta_{2\beta}\left(h_2(X)\delta\right), \; \beta \in \mathscr{B}, \tag{23}$$

where $h_2(X)$ is some $K_2 \times 1$ vector of transformations $h_2(X)$ with regression coefficient δ.

Assumption 10 (Correct Parametric Specification). (i) For each $j = 1, \ldots, J$, there exists a $\gamma_{j,0} \in \mathbb{R}^{K_1}$ such that $p_{j,0}(X) = \zeta_{1j}\left(h_1(X)\gamma_{j,0}\right)$ a.s.; (ii) there exists a $\delta_0 \in \mathbb{R}^{K_2}$ such that for all $\beta \in \mathscr{B}$, $q_0(X, \beta) = \zeta_{2\beta}\left(h_2(X)\delta_0\right)$ a.s.

Assumption 10i holds if the propensity score working model is correctly specified. This restriction is more stringent than in the usual missing data case: the model must be correctly specified for all strata. Assumption 10ii requires the working model for the conditional expectation function to be correctly specified for all β. This is a standard requirement in the analysis of parametric doubly robust estimators.

Assumption 11. (i) For each $j = 1, \ldots, J$, there exists an estimator $\widehat{\gamma}_{j,n}$ such that $\widehat{\gamma}_{j,n} \xrightarrow{p} \gamma_0$ and $\sqrt{n}\left(\widehat{\gamma}_{j,n} - \gamma_0\right) \xrightarrow{d} \mathscr{N}\left(0, \Omega_{\gamma,j}\right)$ and (ii) there exists an estimator $\widehat{\delta}_n$ such that $\widehat{\delta}_n \xrightarrow{p} \delta_0$ and $\sqrt{n}\left(\widehat{\delta}_n - \delta_0\right) \xrightarrow{d} \mathscr{N}(0, \Omega_\delta)$.

Assumption 11 requires that estimators are available that are consistent and asymptotically normal at the parametric rate. This is not a restrictive assumption: the parameters in the working models can typically be estimated using maximum likelihood (for $\gamma_{j,0}$) and nonlinear least squares (for δ_0).

Consider the criterion function

$$\begin{aligned} &G_n^{DR}(\beta) \\ &= \sum_j A_{j,n} \frac{1}{n} \sum_{i=1}^n \left(\frac{s_{ij} d_j \psi(Z_i, \beta)}{\zeta_{1j}\left(h_1(X_i)\widehat{\gamma}_{j,n}\right)} \right. \\ &\quad \left. - \frac{s_{ij} - \zeta_{1j}\left(h_1(X_i)\widehat{\gamma}_{j,n}\right)}{\zeta_{1j}\left(h_1(X_i)\widehat{\gamma}_{j,n}\right)} d_j \zeta_{2\beta}\left(h_2(X_i)\widehat{\delta}_n\right) \right), \end{aligned} \tag{24}$$

and define the *DR* estimator $\widetilde{\beta}_n$ through $G_n^{DR}\left(\widetilde{\beta}_n\right) = 0$. On top of inverse propensity score weighting, the DR estimator

[13] There is a large literature on doubly robust estimation. Some contributions closely related to current setup include Cattaneo (2010), Tan (2010), Graham (2011), Graham et al. (2012, 2016), and Rothe and Firpo (2019).

makes a covariate adjustment based on an estimate of the conditional expectation function.

Assumption 12. (i) The class of functions $\{\zeta_{2\beta}(h_2(\cdot)\delta_0), \beta \in \mathscr{B}\}$ is Glivenko-Cantelli and

$$E\left[\sup_{\beta\in\mathscr{B}}\left\|\zeta_{2\beta}\left(h_2\left(X\right)\delta_0\right)\right\|\right] < \infty;$$

(ii) there exists a $\widetilde{\kappa} > 0$ such that $\zeta_{1j}\left(X\gamma_{j,0}\right) \geq \widetilde{\kappa}$ for all j.

These assumptions guarantee that equation (24) converges uniformly to its limit uniformly. Given that the researcher is in control of the working models, this is not a restrictive assumption.

Assumption 13. (i) For each $j = 1, \ldots, J$, the link function $\zeta_{1j}(\cdot)$ has a derivative ζ'_{1j}, and there exists an $\epsilon_1 > 0$ such that $\sup_{\|\gamma_j-\gamma_{j0}\|<\epsilon_1} E[\|\zeta'_{1j}\left(h_1\left(X\right)\gamma_j\right)h_1\left(X\right)\|] < \infty$; (ii) for each $\beta \in \mathscr{B}$, there exists an $\epsilon_2 > 0$ such that $\sup_{\|\delta-\delta_0\|<\epsilon_2} E[\|\zeta_{2\beta}\left(h_2\left(X_i\right)\delta\right)\|] < \infty$; (iii) for each $\beta \in \mathscr{B}$, the link function $\zeta_{2\beta}$ has a derivative $\zeta'_{2\beta}$, and there exists an $\epsilon_3 > 0$ such that

$$\sup_{\|\delta-\delta_0\|<\epsilon_3} E\left[\left\|\zeta'_{2\beta}\left(h_2\left(X\right)\delta\right)h_2\left(X\right)\right\|\right] < \infty.$$

Assumption 13 imposes some conditions on the working models that guarantee that the resulting class of criterion functions is well behaved. The smoothness assumptions on the working models are stronger than those for the original moment functions, which is reasonable given that the working models are under the control of the researcher.

Theorem 4 *(DR Consistency). If assumptions 1, 2, 3, 6, 7, and 11 hold and at least one of Assumptions 10i or 10ii holds, then* $\widetilde{\beta}_n \xrightarrow{p} \beta_0$.

Theorem 4 provides conditions under which the DR estimator $\widetilde{\beta}_n$ is consistent. In particular, it shows that $\widetilde{\beta}_n$ is indeed doubly robust: only one of the working models needs to be correct for consistency. For asymptotic normality and inference, I impose some additional structure. The results could be generalized to nonsmooth settings by using techniques like those in theorem 3.

Assumption 14 (Additional Smoothness). There exists a $\delta > 0$ such that (i) $\psi(\cdot, \beta)$ is continuously differentiable with respect to β on $\|\beta - \beta_0\| < \delta$, and (ii) $E\left[\sup_{\|\beta-\beta_0\|<\delta}\|\psi\left(Z, \beta\right)\|\right] < \infty$.

Theorem 5. *If the conditions for theorem 4 are satisfied, and assumption 14 holds, then*

$$\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathscr{N}\left(0, \left(\Gamma_0' A' V_A^{-1} A \Gamma_0\right)^{-1}\right),$$

where V_A *is as in theorem 3.*

This theorem says that the DR estimator with correctly specified parametric propensity score and conditional expectation function obtains the same limiting distribution as the IPW estimator with nonparametric propensity score. If the limiting weight matrices are chosen as in remark 3, then the DR estimator is locally efficient.[14]

D. *Inference*

If the working models for the propensity score and the outcome equation are correctly specified, the limiting distributions of the IPW and DR estimators coincide. Recall that the asymptotic variance is given by $\left(\Gamma_0' A' V_A^{-1} A \Gamma_0\right)^{-1}$, where $V_A = \sum_j A_j \Omega_j A_j'$. Consistent standard errors therefore require consistent estimators for Γ_0 and Ω_j, $j = 1, \ldots, J$.

An appropriate estimator for Γ_0 depends on the specific application. For differentiable moment conditions, an analog estimator may be based on the expression for the derivative.[15] Cattaneo (2010, theorem 7) provides a general approach for smooth moment conditions that could be modified for incompletely observed moments. The estimator in Pakes and Pollard (1989) can be used for nonsmooth cases. In what follows, it is assumed that any such consistent estimator $\widehat{\Gamma}_n$ is available.

Recall that the inversely weighted moment conditions have the variance that we are after:

$$\Omega_j = E\left[\frac{s_j}{p_j^2\left(X\right)}\psi\left(Z, \beta_0\right)\psi\left(Z, \beta_0\right)'\right] \tag{25}$$

$$= E\left[\frac{\Sigma_0\left(X\right)}{p_{j,0}\left(X\right)} + q_0\left(X\right)q_0'\left(X\right)\right]. \tag{26}$$

A natural estimator for Ω_j is therefore

$$\widehat{\Omega}_j = \frac{1}{n}\sum_{i=1}^{n}\frac{s_{ij}}{\widehat{p}_j\left(X_i\right)}d_j\psi\left(Z_i, \beta_n\right)\psi\left(Z_i, \beta_n\right)' d_j,$$

where β_n is a consistent estimator for β_0. The following result clarifies the conditions under which consistent standard errors can be based on $\widehat{\Omega}_j$.

Theorem 6. *If* $\beta_n \xrightarrow{p} \beta_0$, $\left\|\widehat{\Gamma}_n - \Gamma_0\right\| = o_p(1)$, *assumptions 5, 6, and 3 hold, and if (i)* ψ *is continuous at* β_0 *almost surely; and (ii) there exists a* $\delta > 0$ *such that* $E\left[\sup_{\|\beta-\beta_0\|<\delta}\|\psi\left(Z, \beta\right)\|^2\right] < \infty$, *then*

$$\left(\widehat{\Gamma}_n' A_n' \widehat{V}_{A,n}^{-1} A_n \widehat{\Gamma}_n\right)^{-1} \xrightarrow{p} \left(\Gamma_0' A' V_A^{-1} A \Gamma_0\right)^{-1},$$

where $\widehat{V}_{A,n} = \sum_j A_{j,n}\widehat{\Omega}_j A_{j,n}'$ *and* $A_n = \sum_j A_{j,n}$.

[14] It attains the efficiency bound, equation (11), which does not incorporate the knowledge about the parametric models if both parametric models are correctly specified.

[15] As an example, the linear IV example has $\Gamma_0 = -E\begin{bmatrix} W_1 X \\ W_2 X \end{bmatrix}$, which can be estimated consistently by using the framework outlined in this paper or by using an IPW estimator from the stratum with complete data.

In the next section, we investigate the finite sample performance of this estimator. An alternative estimator for Ω_j would use estimate Σ_0 and q_0 jointly from the J strata using the expression in equation (26).

E. Simulation Results

This section summarizes the results from a simulation study, the details of which are in appendix D. The study contains four designs. In three of those four, the DGP satisfies MCAR; the fourth satisfies MAR. Two of the four designs are dynamic panel models (section D.1); the other two are fixed effects binary-choice models (section D.2). Identification holds in no stratum in the first design; it holds in each stratum in the remaining designs.

The first design revisits example 2. This design is interesting because identification fails in each stratum and because no estimator is currently available for this example. I document the performance of the IPW estimator and show that its performance increases as (a) ρ increases and (b) the two cohorts become more different in terms of the information they provide.

The second design revisits example 6, a version of example 2 with an additional time period and cohort. In this version, identification holds for each stratum. This design is interesting because existing estimators are available, which allows us to document the efficiency gain of the proposed estimator relative to existing ones. The IPW estimator dominates the other estimators in terms of performance.

The third and fourth designs revisit the fixed-effects binary choice model in example 3. The third design looks at the MCAR case. We quantify the efficiency loss due to incompleteness and show that it is driven by the variance (the bias is negligible) and is lower for the IPW estimator in this paper than for existing procedures. In design 4, we show that the results go through under MAR and that not controlling for selection can lead to severe bias.

VI. Empirical Illustration

This section revisits the analysis in Topalova and Khandelwal (2011, henceforth TK), who investigate the effect of a trade reform using unbalanced firm-level panel data from India. Their paper provides details regarding data and background. Their analysis provides an ideal test for the incomplete data estimators developed in this paper because their data are very unbalanced. For example, fewer than half of the firms are observed over the entire period, and there are 46 distinct patterns. Appendix E contains two figures that describe the incompleteness of the data in more detail.

This section focuses on efficiency gains from using the estimators developed in this paper. An alternative consideration is selection, because it is theoretically possible that selection plays a role in the unbalancedness of the panel. However, a comparison of columns 3 and 4 in table 4 in TK suggests that this is not the case. Furthermore, experiments with propensity scores that depend on the size and age of a firm did not yield different results. For this reason, I present results under MCAR.

TABLE 1.—RESULTS FOR THE STATIC PANEL MODEL

		Available Case		Complete Case		
		TK (3)	Rep.	TK (4)	Rep.	Efficient
β	Estimate	−0.053	−0.035	−0.059	−0.031	−0.043
	SE	(0.016)	(0.013)	(0.017)	(0.011)	(0.009)
	n	14,808	14,808	8,059	8,059	—

Rep.: Replications of results.

Among other contributions, TK estimate the effects of industry-specific output tariffs on the total factor productivity of firms. Their estimates in table 4, columns 3 and 4, are based on a static panel model,

$$pr_{ijt} = \alpha_i + \beta trade_{j,t-1} + X_{ijt}\gamma + u_{ijt}, \tag{27}$$

where i indicates one of 3,108 firms, j indicates a four-digit NIC industry, and t indicates a year in the period 1990 to 1996. The dependent variable *pr* is a productivity measure constructed from production function estimates (see TK, 998). The main explanatory variable *trade* is output tariff measured at the industry level, lagged by one year. In my reexamination, X_{ijt} consists only of time dummies. The quadratic *age* term in TK is closely approximated by the combination of firm and time fixed effects, so that omitting it has almost no effect on the estimated effect of tariffs.

TK estimate this model using the panel fixed-effects (FE) estimator. I instead use a first-difference (FD) estimator. In the static model, FD and FE give very similar results. My reason for using FD is that the estimator for the dynamic model is also estimated via FD.

The FD moment conditions are

$$E\begin{bmatrix} \Delta trade_{j,89}\Delta u_{ij,90} \\ \Delta u_{ij,90} \\ \vdots \\ \Delta trade_{j,95}\Delta u_{ij,96} \\ \Delta u_{ij,96} \end{bmatrix} = 0, \tag{28}$$

where $\Delta u_{ijt} = u_{ijt} - u_{ij,t-1}$, and so on. The moment conditions involving $\Delta trade$ follow from the exogeneity of tariff changes to firm-level decisions. The remaining moment conditions define the time dummies. The incomplete data pattern for a completely observed firm is $D_1 = I_{14}$. A firm that drops out in 1991 has $D_2 = e_{1,2} \otimes I_7$, and so on. In what follows, I discard patterns with fewer than twenty firms. For such patterns, the number of observations is insufficient for the estimation of the corresponding part of the optimal weight matrix.

Table 1 contains the results.[16] The column "TK (3)" reprints the results from column 3 in table 4 of TK. It

[16] Additional descriptive statistics are in appendix E, table 8.

TABLE 2.—RESULTS FOR THE DYNAMIC PANEL MODEL

		TK (6)	Rep.	Efficient
β	Estimate	−0.048	−0.041	−0.037
	SE	(0.013)	(0.016)	(0.013)
ρ	Estimate	0.455	0.472	0.228
	SE	(0.068)	(0.057)	(0.032)

Rep.: Replications of results.

corresponds to an "available case" estimator, which replaces unobservable moment functions by 0s. Column "TK (4)" corresponds to column 4 in table 4 of TK, which implements a complete case estimator, which uses only firms for which all measurements are available in all time periods (a balanced subpanel). The adjacent "Rep" columns contain my replications of those results. The replicated results are slightly different, because (a) I use optimally weighted FD instead of FE; (b) I use bootstrap standard errors rather than robust standard errors; (c) I did not include *age* and age^2 as control variables.

The column "Efficient" implements the estimator proposed in this paper. The main takeaway from table 1 is that this leads to the lowest standard errors, demonstrating the efficiency gains that can be obtained. Table 8 (appendix E) shows that relative efficiency varies with the parameter of interest. However, the estimator proposed in this paper dominates the complete case and available case estimators.

The dynamic model (TK, column 6) adds an autoregressive term to the productivity equation:

$$pr_{ijt} = \alpha_i + \rho\, pr_{ij,t-1} + \beta\, trade_{j,t-1} + X_{ijt}\gamma + \upsilon_{ijt}. \quad (29)$$

TK estimate the parameters in this model using the procedure in Arellano and Bond (1991), which is a GMM estimator based on the moment conditions

$$E\begin{bmatrix} \Delta\, trade_{j,t-1}\Delta\upsilon_{ijt} \\ \Delta\upsilon_{ijt} \\ pr_{ijt-s}\Delta\upsilon_{ijt} \end{bmatrix} = 0, \text{ for } t = 90, \ldots, 96,$$
$$s = t-2, t-3, \ldots, 90.$$

Table 2 compares the result in TK with my replication, as well as with the efficient estimator proposed in this paper. The replication is not exact because my replication does not include the *age* term, as mentioned above. Note that the efficient estimator yields substantial improvements for efficient estimation of the autoregressive parameter (more than 50% relative to TK).

VII. Conclusion

Many data sets used in applied econometrics are incomplete: different information is available for different sampling units. In this paper, I propose a framework for parameter estimation with incomplete data by deriving moment conditions for the incomplete data that generalize those for standard MAR setup for missing data. First, I state conditions under which identification can be obtained with incomplete data. Second, I derive the efficiency bound for this framework. Third, I propose and analyze IPW and DR estimators that attain the efficiency bound. The results are useful for analyzing unbalanced panels, as shown in an application to the analysis of the effect of trade reforms on firm productivity in India.

REFERENCES

Abrevaya, Jason, "Missing Dependent Variables in Fixed-Effects Models," *Journal of Econometrics* 211 (2019), 151–165.

Abrevaya, Jason, and Stephen Donald, "A GMM Approach for Dealing with Missing Data on Regressors and Instruments," Department of Economics, University of Texas at Austin mimeo (2011).

Abrevaya, Jason, and Stephen Donald, "A GMM Approach for Dealing with Missing Data on Regressors," this REVIEW 99 (2017), 657–662.

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo, and James A. Robinson, "Democracy, Redistribution and Inequality" (pp. 1885–1966), in A. B. Atkinson and F. Bourguignon, eds., *Handbook of Income Distribution*, vol. 2 (Amsterdam: Elsevier, 2015).

——— "Democracy Does Cause Growth," *Journal of Political Economy* 127 (2018), 47–100.

Angrist, Joshua D., Victor Lavy, and Analia Schlosser, "Multiple Experiments for the Causal Link between the Quantity and Quality of Children," *Journal of Labor Economics* 28 (2010), 773–824.

Arellano, Manuel, and Stephen Bond, "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies* 58 (1991), 277–297.

Becker Sascha O., and Ludger Woessmann, "Not the Opium of the People: Income and Secularization in a Panel of Prussian Counties," *American Economic Review* 103 (2013), 539–544.

Bickel, Peter J., Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models* (New York: Springer, 1993).

Cameron, A. Colin, and Pravin K. Trivedi, *Microeconometrics: Methods and Applications* (Cambridge: Cambridge University Press, 2005).

Card, David, "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in L. Christofides, E. K. Grant, and R. Swindinsky, eds., *Aspects of Labour Economics: Essays in Honour of John Vanderkamp* (Toronto: University of Toronto Press, 1995).

Cattaneo, Matias D., "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects under Ignorability," *Journal of Econometrics* 155 (2010), 138–154.

Chamberlain, Gary, "Analysis of Covariance with Qualitative Data," *Review of Economic Studies* 47 (1980), 225–238.

Chamberlain, Gary, "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics* 34 (1987), 305–334.

——— "Efficiency Bounds for Semiparametric Regression," *Econometrica* 60 (1992a), 567–596.

——— "Comment: Sequential Moment Restrictions in Panel Data," *Journal of Business and Economic Statistics* 10 (1992b), 20–26.

Chaudhuri, Saraswata, and David K. Guilkey, "GMM with Multiple Missing Variables," *Journal of Applied Econometrics* 31 (2016), 678–706.

Chen, Baojiang, Grace Yi, and Richard J. Cook, "Weighted Generalized Estimating Functions for Longitudinal Response and Covariate Data That Are Missing at Random," *Journal of the American Statistical Association* 105 (2010), 336–353.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi, "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics* 36 (2008), 808–843.

Dagenais, M. G., "The Use of Incomplete Observations in Multiple Regression Analysis: A Generalized Least Squares Approach," *Journal of Econometrics* 1 (1973), 317–328.

Dardanoni, Valentino, Salvatore Modica, and Franco Peracchi, "Regression with Imputed Covariates: A Generalized Missing-Indicator Approach," *Journal of Econometrics* 162 (2011), 362–368.

de Loecker, Jan, and Frederic Warzynski, "Markups and Firm-Level Export Status," *American Economic Review* 102 (2012), 2437–2471.
Feng, Qian, "Essays in Causal Inference with Endogeneity and Missing Data," PhD dissertation, University of Texas at Austin (2018).
Gourieroux, Christian, and Alain Monfort, "On the Problem of Missing Data in Linear Models," *Review of Economic Studies* 48 (1981), 579–586.
Graham, Bryan S., "Efficiency Bounds for Missing Data Models with Semiparametric Restrictions," *Econometrica* 79 (2011), 437–452.
Graham, Bryan S., Cristine Campos de Xavier Pinto, and Daniel Egel, "Inverse Probability Tilting for Moment Condition Models with Missing Data," *Review of Economic Studies* 79 (2012), 1053–1079.
——— "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)," *Journal of Business and Economic Statistics* 34 (2016), 288–301.
Hahn, J., "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (1998), 315–331.
Hirano, Keisuke, Guido Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71 (2003), 1161–1189.
Hirano, Keisuke, Guido Imbens, Geert Ridder, and Donald B. Rubin, "Combining Panel Data Sets with Attrition and Refreshment Samples," *Econometrica* 69 (2001), 1645–1659.
Hristache, Marian, and Valentin Patilea, "Semiparametric Efficiency Bounds for Conditional Moment Restriction Models with Different Conditioning Variables," *Econometric Theory* 32 (2016), 917–946.
Hristache, Marian, and Valentin Patilea, "Conditional Moment Models with Data Missing at Random," *Biometrika* 104 (2017), 735–742.
Mogstad, Magne, and Matthew Wiswall, "Instrumental Variables Estimation with Partially Missing Instruments," *Economics Letters* 114 (2012), 186–189.
Pacini, David, and Frank Windmeijer, "Moment Conditions for AR(1) Panel Data Models with Missing Outcomes," *Bristol Economics discussion papers* 15/660 (2015).
Pakes, Ariel, and David Pollard, "Simulation and the Asymptotics of Optimization Estimators," *Econometrica* 57 (1989), 1027–1057.
Papke, Leslie E., "The Effects of Spending on Test Pass Rates: Evidence from Michigan," *Journal of Public Economics* 89 (2005), pp. 821–839.
Papke, Leslie E., and Jeffrey M. Wooldridge, "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates," *Journal of Econometrics* 145 (2008), 121–133.
Prokhorov, Artem, and Peter Schmidt, "GMM Redundancy Results for General Missing Data Problems," *Journal of Econometrics* 151 (2009), 47–55.
Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association* 89 (1994), 846–866.
Rodrik, Dani, Arvind Subramanian, and Francesco Trebbi, "Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development," *Journal of Economic Growth* 9 (2004), 131–165.
Rothe, Christoph, and Sergio Firpo, "Properties of Doubly Robust Estimators When Nuisance Functions Are Estimated Nonparametrically," *Econometric Theory* 35 (2019), 1048–1087.
Schularick, Moritz, and Thomas M. Steger, "Financial Integration, Investment, and Economic Growth: Evidence from Two Eras of Financial Globalization," this REVIEW 92 (2010), 756–768.
Sturm, Jan-Egbert, and Jakob de Haan, "Income Inequality, Capitalism, and Ethno-Linguistic Fractionalization," *American Economic Review* 105 (2015), 593–597.
Tan, Zhiqiang, "Bounded, Efficient, and Doubly Robust Estimation with Inverse Weighting," *Biometrika* 97 (2010), 661–682.
Topalova, Petia, and Amit Khandelwal, "Trade Liberalization and Firm Productivity: The Case of India," this REVIEW 93 (2011), 995–1009.
Tsiatis, Anastasios A., *Semiparametric Theory and Missing Data* (New York: Springer, 2010).
Verbeek, Marno, and Theo Nijman, "Testing for Selectivity Bias in Panel Data Models," *International Economic Review* 33 (1992), 681–703.
Wooldridge, Jeffrey M., "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics* 141 (2007), 1281–1301.
Yagan, Danny, "Capital Tax Reform and the Real Economy: The Effects of the 2003 Dividend Tax Cut," *American Economic Review* 105 (2015), 3531–3563.

This article has been cited by:

1. Pedro H. C. Sant'Anna, Jun B. Zhao. 2018. Doubly Robust Difference-in-Differences Estimators. *SSRN Electronic Journal* . [Crossref]
2. Yiqun Ma. 2015. Uncertainty and investment: Evidence from the Australian mining industry. *Resources Policy* **46**, 191-201. [Crossref]